Node Feature characteristics: Content-based: Natural language description (cb_nld) and Content-based: Literals (cb_Literals)

For the comparison, node features are classified into three categories: Content-based NLD (indicated as NLD), Content-based other Literals (indicated as Literals\NLD), and Topology-based (indicated as Topology). Figure 1 illustrates the distinction of the different feature characteristics. Thereby, content-based NLD refers to encoded NLD, while Content-based Literals\NLD refer to encoded content attributes excluding NLD (such as numeric, categorical or boolean values). Topology-based features refer to features that encode the topological structure of the graph.

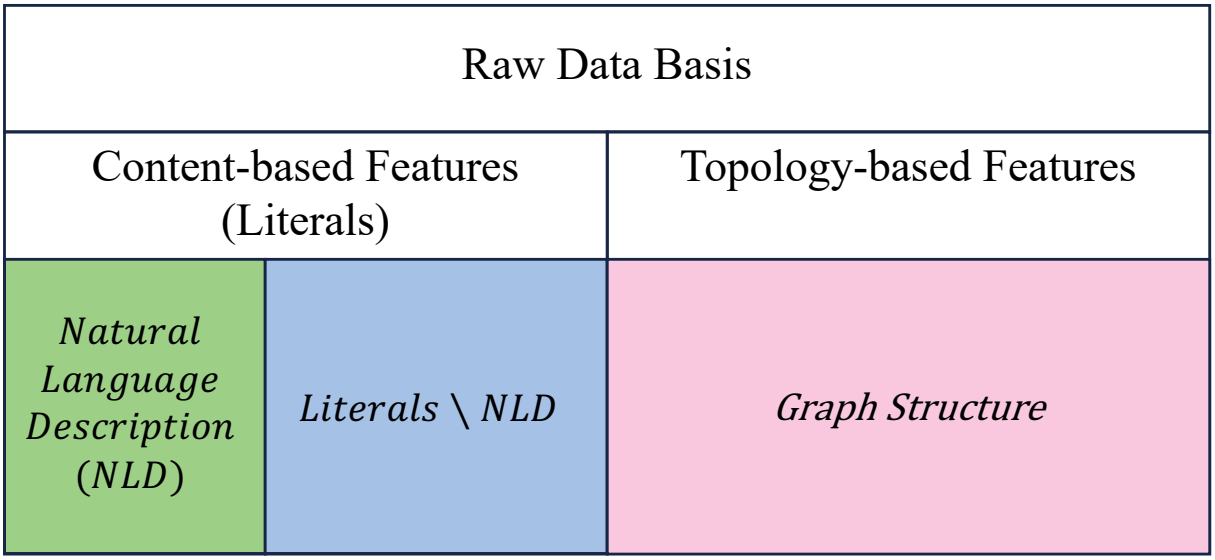

Figure 1: Different node feature characteristics.

SemOpenAlex-SemanticWeb:

For **SOA-SW** the following data type properties are used to calculate the content-based node features (cb_literals):

- **Work (out of 18 data type properties, 8 are selected):** `dcterms:created, dcterms:modified, fabio:hasPublicationYear, soa:citedByCount, soa:cross- refType, soa:isRetracted, dcterms:title, and dcterms:abstract.`
- **Author (out of 10 data type properties, 4 are selected):** `dcterms:created, dcterms:modified, soa:citedByCount, and soa:worksCount.`
- **Concept (out of 11 data type properties, 4 are selected):** `dcterms:modified, soa:worksCount, dcterms:created, and soa:level.`
- **Source (out of 19 data type properties, 10 are selected):** `dcterms:created, dcterms:modified, bido:h-index, gn:countryCode, soa:2YrMeanCitedness, soa:i10Index, soa:worksCount, soa:isInDoaj, soa:isOa, and soa:sourceType.`

- **Institution (out of 14 data type properties, 5 are selected):** `dcterms:created`, `dcterms:modified`, `dbp:countryCode`, `soa:worksCount`, and `soa:rorType`.
- **Publisher (out of 12 data type properties, 5 are selected):** `dcterms:modified`, `bido:h-index`, `gn:country-code`, `soa:i10Index`, and `soa:level`.

AutoRDF2GML detects NLD features (cb_nld) for the work nodes, specifically the properties dcterms:title and dcterms:abstract. The work titles and abstracts (`dcterms:title` and `dcterms:abstract`) are concatenated, and subsequently, a 128-dimensional embedding is generated for this combined data using SciBert.

LPWC:

For **LPWC** the following data type properties are used to calculate the content-based node features (cb_literals):

• **Paper (out of 7 data type properties, 3 are selected):** `dcterms:date`, `dcterms:abstract`, and `dcterms:title`.
• **Method (out of 7 data type properties, 4 are selected):** `dcterms:description`, `dbp:fullname`, `lpwc:numberPapers`, and `lpwc:introducedYear`.
• **Task (out of 2 data type properties, 2 are selected):** `dcterms:description`, and `foaf:name`.
• **Dataset (out of 10 data type properties, 7 are selected):** `dcterms:description`, `dcterms:title`, `dbp:fullname`, `dcterms:issued`, `dcterms:language`, `lpwc:modality`, and `lpwc:numberPapers`.

AutoRDF2GML detects NLD features (cb_nld) for all node types, specifically:
• For **Paper** nodes `dcterms:title`, and `dcterms:abstract`.
• For **Method** nodes `dbp:fullname`, and `dcterms:description`.
• For **Task** nodes `foaf:name`, and `dcterms:description`.
• For **Dataset** nodes `dcterms:title`, `dbp:fullname`, and `dcterms:description`.
The detected NLD features are concatenated, and then a 128-dimensional embedding is calculated for the combined data using SciBert.